Semi-supervised siamese network using self-supervision under scarce annotation improves class separability and robustness to attack

Gabriel B. Cavallari and Moacir A. Ponti
ICMC – Universidade de São Paulo, São Carlos-SP, Brazil
Email: gabriel.cavallari@usp.br, ponti@usp.br

***Abstract*—Self-supervised learning approaches were shown to benefit feature learning by training models under a pretext task. In this context, learning from limited data can be tackled using a combination of semi-supervised learning and self-supervision. In this paper we combine the traditional supervised learning paradigm with the rotation prediction self-supervised task, that are used simultaneously to train a siamese model with a joint loss function and shared weights. In particular, we are interested in the case in which the proportion of labeled with respect to unlabeled data is small. We investigate the effectiveness of a compact feature space obtained after training under such limited annotation scenario, in terms of linear class separability and under attack. The study includes images from multiple domains, such as natural images (STL-10 dataset), products (Fashion-MNIST dataset) and biomedical images (Malaria dataset). We show that in scenarios where we have only a few labeled data the model augmented with a self-supervised task can take advantage of the unlabeled data to improve the learned representation in terms of the linear discrimination, as well as allowing learning even under attack. Also, we discuss the choices in terms of self-supervision and cases of failure considering the different datasets.**

I. Introduction

Deep convolutional neural networks have been successful in computer vision thanks to their ability to learn high-level semantic visual representations, which have enabled remarkable performance in various tasks [1]–[3]. Current computer vision systems demonstrate excellent performance in a variety of benchmarks, such as object detection, image recognition and semantic segmentation. These networks mainly follow the supervised learning paradigm, in which many input-output pairs are required for training. However, large amounts of manually labeled data are costly, time-consuming, complex and expensive to obtain, and some real-world applications require categories that are not present in standard large-scale benchmark datasets. Therefore, investigate strategies to learn without or under limited annotated data [4], [5] is of great importance to take advantage of the large availability of unsupervised data.

Recently, self-supervised learning methods [6] demonstrated promising results using only unlabeled data, being able to learn features that are competitive with respect to carefully tuned supervised baselines [7]–[10]. Those methods are trained to solve pretext tasks that require high-level semantic understating to be solved, producing useful representations that can be used for solving other downstream tasks such as image recognition.

Semi-supervised learning [11] allow using unlabeled data in deep learning [5], and since self-supervised learning can leverage unlabeled data and mitigate the hunger of deep networks, it has potential to improve the final model. Indeed, state-of-the art semi-supervised models are able to perform at the same level as strong supervised models while using only a fraction of the labeled data [12]–[17].

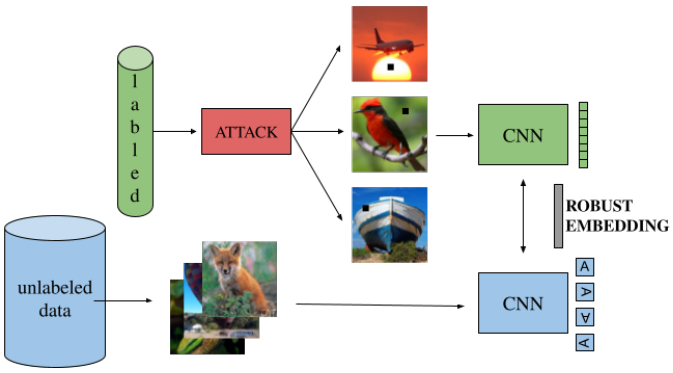


Fig. 1. An illustration of our approach: while labeled images can be used to train a classification-based CNN (in green), unlabeled data is used to learn an auxiliary task, producing an embedding that is more robust to attack and more discriminative. The pixel attack is deliberately exaggerated in this image for the sake of illustration.

Multiple studies explore self-supervised tasks and semi-supervised methods with the purpose of learning good features with less human-annotated data. Usually they test the learned features by fine-tuning the model on other tasks, but it remains uncertain if the standard evaluation protocol is sufficiently strong [18]. Introducing noise to the training data may also be useful to study the model's robustness [19], as well as designing robust models [20]. However, understanding different aspects of those models when trained on distinct image domains has not yet been fully investigated, specially in scenarios where noise comes from the training data.

We hypothesize that a self-supervised task, even a simple one like the rotation-prediction task, can help the process of learning image representations in a scenario that we have both few labeled data and under 1-pixel attacks as noise. In this

work, by using a setup as illustrated in Figure 1, we compare the effectiveness of this method on 3 different image domains, also studying the robustness of the method to noise in the form of 1-pixel attacks. We show it is possible to use unlabeled data to obtain more discriminative and more robust representations compared to the supervised baselines when evaluating the resulting feature spaces.

II. Related Work

In this section we review the most relevant developments in the fields related to our work, and summarize two particular categories most related to this work: self-supervised learning and consistency regularization.

A. Self-supervised learning

Self-supervised learning relies on pretext tasks that can be formulated using only unsupervised data. By producing surrogate labels, those tasks make use of those generated labels to guide the learning process. These models learn useful image representations in order to solve those tasks and achieve state-of-the-art performance when we consider methods that rely only on unlabeled images. Auxiliary tasks were shown to be important for the context of minimal data-learning [21] and to allow, in the context of images, to learn low-level features as good as via strong supervision [22].

One of the first methods of self-supervision was based on patches generated from the image. In [23], the authors train a CNN model that predicts relative location of two randomly sampled non-overlapping image patches. Another method solves jigsaw puzzles considering image tiles obtained from the image [24]. Also, clustering techniques were proposed to improve self-supervised learning [25]. Some self-supervised techniques employ image-level losses, instead of using image patches. Grayscale image colorization task was proposed as a pretext task [26]. It is also possible to learn useful features by predicting simple rotation transformations [27].

An important step towards good self-supervised tasks was to encourage models to learn representations that are invariant to heavy image augmentations, by also imposing restrictions on the representation space learned [28]. It is also possible to consider what [29] calls as visual primitives, requiring that the sum of representations of all image patches should be close to the representation of the whole image. K-means clustering can also be used to produce pseudo-labels for the data as presented in [30], which was one of the first works to accomplish competitive performance to supervised models, specifically AlexNet.

The use of contrastive learning became popular more recently [31]. SWAV [10] proposes a swapped prediction contrastive objective with online clustering to deal with multi-view augmentation. SimCLR [8] learns representations by maximizing agreement between differently augmented views of the same image in the latent space. MoCo [7] maintains a queue of negative samples and uses a moving-averaged encoder to improve the queue consistency. Those studies achieve competitive results on the ImageNet dataset compared to the supervised baselines.

B. Consistency Regularization

Semi-supervised learning (SSL) is a class of algorithms that learn considering both labeled and unlabeled data. Consistency regularization methods add auxiliary loss terms computed on the unlabeled data. The auxiliary loss terms can be considered as a regularizer. π-Model [13], Mean Teacher [12] and Virtual Adversarial Training [14] that take advantage of consistency losses, among other works.

In terms of semi-supervised training strategy and the use of the rotation-prediction as an auxiliary task, our work is most related to S4L [32], SESEMI [33] and [34]. In [32] the authors train semi-supervised models with the rotation-prediction auxiliary and also other tasks, on the ImageNet dataset. In [33], the author uses the rotation-prediction task as an auxiliary loss term to train the model on SVHN, CIFAR-10 and CIFAR-100. Different from those two, we test the method not only in natural and color images, but also in a grayscale (Fashion-MNIST) image domain and also the Malaria dataset, a dataset that is not angle oriented (unlike photographs that have angle bias). In [34] the authors find that self-supervision can increase model's robustness to adversarial examples, label and input corruptions. They use the rotation-prediction task as an auxiliary loss term to train the semi-supervised model on CIFAR-10. In our work we perform 1-pixel attacks and evaluate the resulting feature representations.

III. Method

In this section we present the general description of our approach. We do not intent to compare our results with state-of-the art semi-supervised classification methods, but rather to evaluate how discriminative (and robust) are the representations obtained from the different methods and strategies. In particular we are interested in the semi-supervised image classification problem of different domains, including natural and biomedical images.

Let a dataset D_l containing N_l pairs of images and labels, and an unlabeled dataset D_u that contains N_u images without annotation. Our semi-supervised method considers a **joint classification loss** L to train a siamese network:

$$\mathcal{L}_{SS} = \lambda_l \cdot \ell_l(D_l) + \lambda_u \cdot \ell_u(D_u) \tag{1}$$

where both ℓ_l and ℓ_u optimize a cross-entropy loss function: the former with a supervised paradigm, and the latter with a self-supervised task. The weights λ_l and λ_u are non-negative numbers. The loss function $\mathcal{L}$ can be used under different self-supervised losses ℓ_u.

In this work we focus on the rotation prediction self-supervised task. In this task the network must predict one of the four rotation degrees (0º, 90º, 180º, 270º) applied to the image, turning the task into a 4-class classification problem.

Our siamese network has shared weights and receives mini-batches containing an equal amount of labeled and unlabeled images (balancing the supervised and self-supervised tasks).

Algorithm 1 Pseudocode, Keras-like style

```
#load base network (pre-softmax)
base_encoder = base_network()

input_lab = Input(name='input_lab')
input_rot = Input(name='input_rot')

encoder_labeled = base_encoder(input_lab)
encoder_rotation = base_encoder(input_rot)

y_labeled = Dense(n_classes, 'softmax',
            name='y_lab')(encoder_labeled)
y_rotation = Dense(4, 'softmax',
            name='y_rot')(encoder_rotation)

semi = Model(inputs=[input_lab, input_rot],
            outputs=[y_labeled, y_rotation])

semi.compile(optimizer=opt,
  loss={'y_lab':'categorical_crossentropy',
        'y_rot':'categorical_crossentropy'})

# train_generator is a generator that
# returns [x_lab, x_rot], [y_lab, y_rot]
history = semi.fit(train_generator)
```

At each epoch, the model sees all unlabeled images N_u, while the labeled images N_l are seen by the model a total of N_u/N_l times. Because $N_u > N_l$ and we used balanced batches, that way we can make sure that in one epoch the model sees all the unlabeled image set. Therefore, our network will see more, although repeated, labeled instances per epoch when compared to the fully supervised one. We compensate that by allowing more epochs for the supervised setting, until convergence.

In Figure 2 we illustrate the semi-supervised network, while Algorithm 1 shows the pseudocode for our network.

A. Datasets

We assess the performance of our method on three different datasets: STL-10 [35], Fashion-MNIST [36] and Malaria [37]. STL-10 is an image recognition dataset designed for semi-supervised and unsupervised feature learning, containing 96×96 color images of airplanes, birds, cars, cats, deers, dogs, horses, monkeys, ships and trucks. Fashion-MNIST has 28×28 grayscale images with centered pieces of clothing and fashion accessories. Malaria dataset contains cell images of multiple resolutions with instances of parasitized and uninfected cells from the thin blood smear slide images of segmented cells. Figure 3 shows examples of images from those datasets.

B. Experimental setup

The following CNN backbones were investigated: MobilenetV2, InceptionV3 and ResNet50v2. For all backbones, before the prediction layers, we added 2 fully connected layers of size 4096 and 128, both with relu activation. The final representation is obtained from the 128-D layer.

On both supervised baseline and the semi-supervised approach, we used a fixed learning rate of 0.001 with an exponential learning rate decay starting at 2/5 of total epochs. For all experiments reported in this paper, we used $\lambda_l = 1.0$ and $\lambda_u = 1.0$.

We employed a batch size of 64 and Adam optimizer. We do not use data augmentation. The supervised baselines for all datasets were trained for 100 epochs. The semi-supervised baselines were trained for 30, 50 and 100 epochs for the Fashion-MNIST, STL-10 and Malaria datasets, respectively. We train each model 5 times, each time with a different random split of labeled data. The random labeled images sets used in the supervised baseline are the same used in the semi-supervised model. All models, supervised and semi-supervised, were trained from scratch.

For STL-10, we maintained the original image size of 96×96. For Fashion-MNIST, images were resized to 96×96. For Malaria dataset, we resized the images to 128×128 .

C. Evaluation

We investigate a scenario of limited availability of labels under supervised and semi-supervised learning. In particular, all experiments will consider a small fraction (1% and 5%) of the labeled data with respect to the total amount of unlabeled data N_u, i.e. N_l will be either $0.01N_u$ or $0.05N_u$. A supervised method using this fraction is used as the baseline of our experiments. Under semi-supervised mode, the same small fraction of the labeled data is used, but with addition of unlabeled data.

The 1-pixel attack case was produced assuming access to the training data. Thus, we attack images contained in D_l. For each class we arbitrarily choose a pixel location where the value is always the same on every image of that class as illustrated in Figure 4.

In order to evaluate the discriminative capability of the learned representations, we employed a linear SVM on the 128-D extracted features. The SVM classifier is a shallow classifier with low complexity bias and low sensitivity to parameter tuning, also having strong learning guarantees that make it useful as a tool to evaluate linear separability of feature spaces [38]. After extracting features from both the training and testing sets using the network model, the SVM is trained using the features obtained from the training set, i.e. the 1% or 5% fraction of the original data, using no kernels and parameter $C = 1$. Then, the features obtained from the test set (never seen during network training or SVM training) are used to obtain the accuracies reported as result.

IV. Results

The STL-10 dataset originally consists of 100,000 images in the unlabeled set, 5,000 images in the training set and 8,000 images in the test set. When training the supervised baselines, we use either 1,000 or 5,000 images from the training set as labeled images (1% or 5% of images in relation to the total unsupervised set). When training the semi-supervised model, we use same fractions of labeled data but also use the whole 100,000 unlabeled images. The results for the STL-10 dataset are shown in Table I for 1% of labelled data and in Table II for

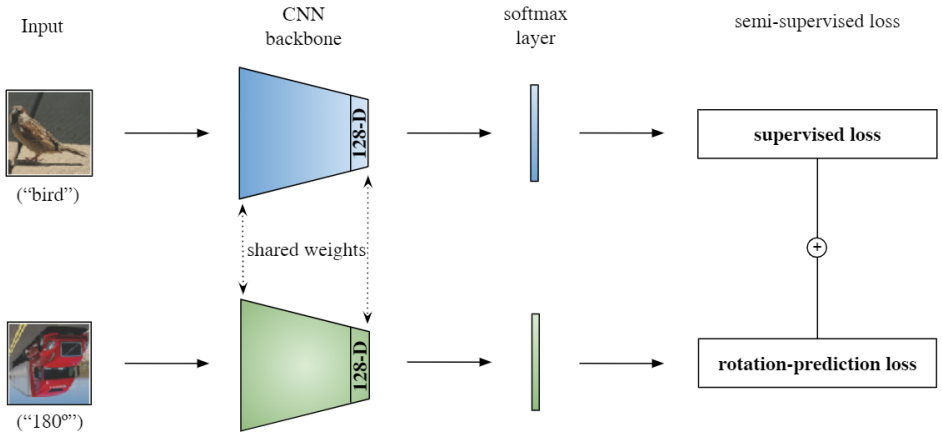


Fig. 2. An illustration of the method for semi-supervised learning. The CNN has shared weights so that both the classification and the auxiliary task play a role in learning features, which is guided by the semi-supervised loss.

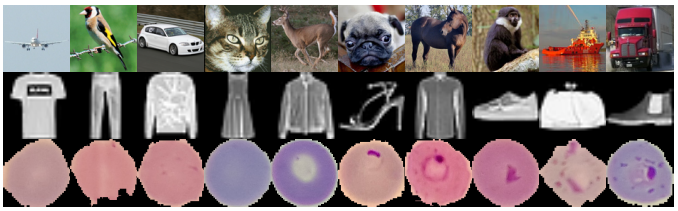


Fig. 3. Examples of STL-10, Fashion-MNIST and Malaria image classes, respectively.

TABLE I
LINEAR SVM TEST ACCURARY OF THE STL-10 DATASET WHEN USING 1% OF LABELED DATA.

Architecture	Supervised (1% of data)		Semi-supervised (1% + unlabeled)		Δ	
	Pixel attack	No attack	Pixel attack	No attack	Pixel attack	No attack
Mobilenetv2	10.0 ± 0.0	10.0 ± 0.0	31.9 ± 2.7	44.9 ± 2.7	+21.9	+34.9
Inceptionv3	40.3 ± 3.1	43.4 ± 2.5	38.9 ± 2.4	48.4 ± 3.5	-1.4	+5.0
Resnet50v2	33.9 ± 3.6	34.8 ± 1.3	40.2 ± 5.3	46.6 ± 2.4	+6.3	+11.8

5% of labeled data. In all tables, Δ represents the difference between using only supervised and semi-supervised.

The Fashion-MNIST dataset originally consists of 60,000 images in the training set and 10,000 images in the test set. When training the supervised baselines, we use either 1% (600) or 5% (3,000) of labeled images. When training the semi-supervised model, we use the same fractions of labeled data but also use the whole 60,000 images as unlabeled data. The results for the Fashion-MNIST dataset using 1% and 5% of labeled data are shown, respectively, in Tables III and IV.

Malaria dataset originally consists of 27,558 labeled images in total. For our experiments we consider half of total images for the training set and the other half for the test set. When training the supervised baselines, we use either 1% (137) or 5% (685) of the training set. When training the semi-supervised model, we use the same fractions of labeled data but also use the whole training set of 13,779 images as unlabeled data. The results for the Malaria dataset are shown in Tables V and VI.

TABLE II
LINEAR SVM TEST ACCURARY OF THE STL-10 DATASET WHEN USING 5% OF LABELED DATA

Architecture	Supervised (5% of data)		Semi-supervised (5% + unlabeled)		Δ	
	Pixel attack	No attack	Pixel attack	No attack	Pixel attack	No attack
Mobilenetv2	24.9 ± 3.7	54.3 ± 0.4	37.5 ± 2.2	63.9 ± 3.3	+12.6	+9.6
Inceptionv3	43.6 ± 3.8	64.2 ± 1.8	51.2 ± 3.4	65.6 ± 3.0	+7.6	+1.4
Resnet50v2	34.0 ± 2.7	56.6 ± 2.4	42.3 ± 3.0	64.1 ± 2.2	+8.3	+7.5

TABLE III
LINEAR SVM TEST ACCURARY OF THE FASHION DATASET WHEN USING 1% OF LABELED DATA

Architecture	Supervised (1% of data)		Semi-supervised (1% + unlabeled)		Δ	
	Pixel attack	No attack	Pixel attack	No attack	Pixel attack	No attack
Mobilenetv2	10.0 ± 0	10.0 ± 0	69.2 ± 2.7	79.6 ± 1.9	+59.2	+69.6
Inceptionv3	70.9 ± 2.2	77.7 ± 0.8	67.9 ± 1.8	80.3 ± 1.6	-3.0	+2.6
Resnet50v2	71.1 ± 1.7	77.0 ± 1.9	73.3 ± 0.6	79.3 ± 5.7	+2.2	+2.3

A. *Class precision/recall and Visualization of feature spaces*

In order to understand in more detail the effects of using semi-supervised learning based on a self-supervised auxiliary task, we visualized (using t-SNE projection of the 128-D features from the test set) some scenarios and showed the class-wise precision/recall values.

First, the STL-10 results with 5% of the labeled data, it is possible to see better clusters in both no attack (Figure 5) and attack (Figure 6 see better clusters for classes such as car, bird and truck) visualizations, although the improvement without attack is marginal. From Table VII we can see that the semi-supervised learning helped improving the average recall by a small margin, which may have slightly impacted the accuracy and F1-score. Under attack, as shown in Table VIII, the improvement is more evident in all metrics, decreasing also the recall standard deviation.

For the Fashion-MNIST dataset, we confirm the remarkable improvement on the learned features, when visualizing the

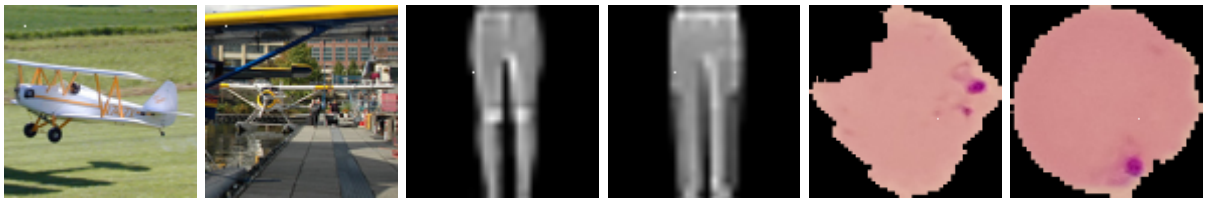


Fig. 4. Examples of 1-pixel attack. From left to right: two images of airplanes from STL-10, two images of trousers from Fashion-MNIST and two images of infected samples from Malaria. Best viewed with zoom.

TABLE IV
LINEAR SVM TEST ACCURARY OF THE FASHION DATASET WHEN USING 5% OF LABELED DATA

Architecture	Supervised (5% of data)		Semi-supervised (5% + unlabeled)		Δ	
	Pixel attack	No attack	Pixel attack	No attack	Pixel attack	No attack
Mobilenetv2	36.2 ± 8.5	48.2 ± 2.1	74.5 ± 2.0	87.3 ± 0.2	+38.3	+39.1
Inceptionv3	66.1 ± 3.5	86.9 ± 0.9	72.4 ± 1.0	87.8 ± 0.6	+6.3	+0.9
Resnet50v2	73.5 ± 0.7	85.9 ± 0.5	77.4 ± 1.7	87.5 ± 0.2	+3.9	+1.6

TABLE V
LINEAR SVM TEST ACCURARY OF THE MALARIA DATASET WHEN USING 1% OF LABELED DATA

Architecture	Supervised (1% of data)		Semi-supervised (1% + unlabeled)		Δ	
	Pixel attack	No attack	Pixel attack	No attack	Pixel attack	No attack
Mobilenetv2	49.9 ± 0	49.9 ± 0	54.8 ± 2.0	88.3 ± 5.0	+4.9	+38.4
Inceptionv3	55.7 ± 7.3	60.2 ± 9.0	53.7 ± 2.4	92.3 ± 1.2	-2.0	+32.1
Resnet50v2	53.8 ± 1.7	54.6 ± 4.3	60.5 ± 3.7	89.0 ± 0.8	+6.7	+34.4

spaces without attack in Figure 7.

Finally, Malaria may be the more interesting case. Note how the improvement again balances out the results from both classes as show in Table IX. Also the space was completely overlapped, and was improved significantly as shown in Figure 8. Even for the attack case, in which the performance was not significantly improved, the visualization shows the space to be better formed (see Figure 9) as it appears to be a gradient from more uninfected samples (in cyan, on the left hand side) to more dense infected samples (in red, on the right hand side of the visualization plane).

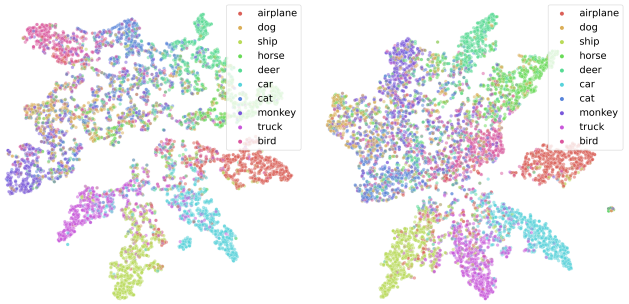


Fig. 5. tSNE visualization for STL-10 5% Inception v3, without attack. Left: supervised, Right: semi-supervised

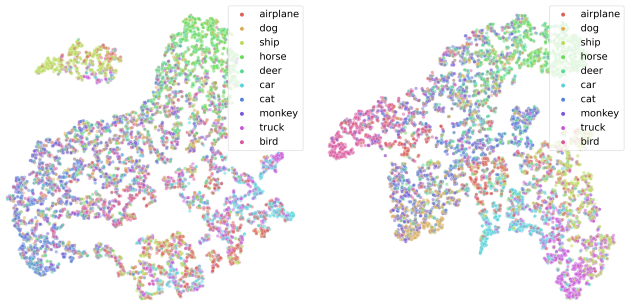


Fig. 6. tSNE visualization for STL-10 5% Inception v3, with attack. Left: supervised, Right: semi-supervised

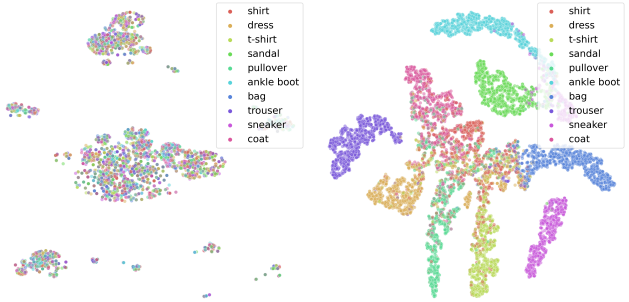


Fig. 7. tSNE visualization for Fashion 1% Mobilenet v2, without attack. Left: supervised, Right: semi-supervised.

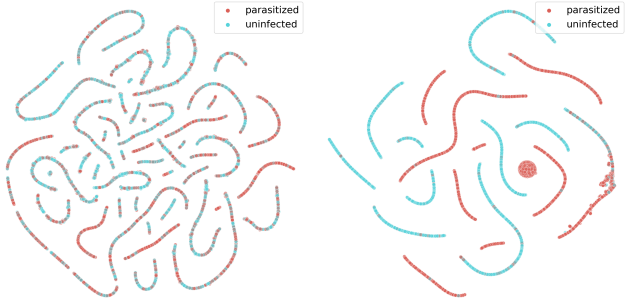


Fig. 8. tSNE visualization for Malaria 1% Inception v3, without attack. Left: supervised, Right: semi-supervised.

TABLE VI
LINEAR SVM TEST ACCURACY OF THE MALARIA DATASET WHEN USING 5% OF LABELED DATA

Architecture	Supervised (5% of data)		Semi-supervised (5% + unlabeled)		Δ	
	Pixel attack	No attack	Pixel attack	No attack	Pixel attack	No attack
Mobilenetv2	49.9 ± 0	49.9 ± 0	57.6 ± 3.8	94.4 ± 0.4	+7.7	+44.5
Inceptionv3	69.2 ± 15.2	94.3 ± 0.4	57.0 ± 1.0	94.3 ± 0.5	-12.2	+0.0
Resnet50v2	83.2 ± 15.0	93.7 ± 0.7	59.2 ± 4.1	93.6 ± 0.7	-24.0	-0.1

TABLE VII
CLASS-WISE PRECISION, RECALL AND F1-SCORE FOR STL-10 5% INCEPTION V3 WITHOUT ATTACK.

Supervised			
class	precision	recall	f1-score
airplane	0.77	0.80	0.78
bird	0.61	0.53	0.57
car	0.79	0.77	0.78
cat	0.44	0.47	0.46
deer	0.63	0.57	0.60
dog	0.40	0.51	0.44
horse	0.66	0.66	0.66
monkey	0.55	0.51	0.53
ship	0.77	0.78	0.77
truck	0.72	0.68	0.70
mean/std	0.63 ± 0.13	0.62 ± 0.12	0.62 ± 0.12
accuracy		0.631	
Semi-Supervised			
class	precision	recall	f1-score
airplane	0.83	0.79	0.81
bird	0.53	0.52	0.53
car	0.85	0.81	0.83
cat	0.42	0.53	0.47
deer	0.64	0.59	0.61
dog	0.43	0.41	0.42
horse	0.65	0.62	0.64
monkey	0.51	0.52	0.52
ship	0.79	0.78	0.78
truck	0.73	0.75	0.74
mean/std	0.63 ± 0.16	0.63 ± 0.14	0.63 ± 0.14
accuracy		0.636	

B. Convergence of the proposed loss

In order to show how the proposed loss compares with a regular classification loss in terms of convergence, we show in Figures 10 and 11 the curves for the classification network loss and for the semi-supervised network. Note that we are able to converge smoothly to a low value in less epochs than

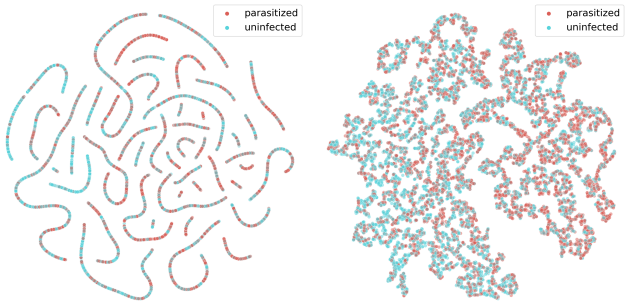


Fig. 9. tSNE visualization for Malaria 1% Inception v3, with attack. Left: supervised, Right: semi-supervised.

TABLE VIII
CLASS-WISE PRECISION, RECALL AND F1-SCORE FOR STL-10 5% INCEPTION V3 WITH ATTACK.

Supervised			
class	precision	recall	f1-score
airplane	0.47	0.50	0.49
bird	0.30	0.27	0.28
car	0.52	0.42	0.46
cat	0.35	0.59	0.44
deer	0.25	0.14	0.18
dog	0.25	0.16	0.20
horse	0.54	0.68	0.61
monkey	0.22	0.21	0.22
ship	0.59	0.65	0.62
truck	0.40	0.41	0.41
mean/std	0.38 ± 0.13	0.40 ± 0.20	0.39 ± 0.16
accuracy		0.407	
Semi-Supervised			
class	precision	recall	f1-score
airplane	0.51	0.49	0.50
bird	0.67	0.60	0.63
car	0.61	0.54	0.57
cat	0.39	0.42	0.40
deer	0.30	0.32	0.31
dog	0.47	0.36	0.41
horse	0.74	0.66	0.69
monkey	0.27	0.33	0.30
ship	0.61	0.61	0.61
truck	0.58	0.69	0.63
mean/std	0.51 ± 0.15	0.50 ± 0.13	0.50 ± 0.14
accuracy		0.506	

TABLE IX
CLASS-WISE PRECISION, RECALL AND F1-SCORE FOR MALARIA 1% INCEPTION V3 WITHOUT ATTACK.

Supervised			
class	precision	recall	f1-score
uninf.	0.58	0.87	0.69
parasit.	0.75	0.37	0.49
mean/std	0.66 ± 0.12	0.62 ± 0.35	0.59 ± 0.14
accuracy		0.624	
Semi-Supervised			
class	precision	recall	f1-score
uninf.	0.91	0.96	0.93
parasit.	0.96	0.90	0.93
mean/std	0.93 ± 0.03	0.93 ± 0.04	0.93 ± 0.0
accuracy		0.937	

TABLE X
CLASS-WISE PRECISION, RECALL AND F1-SCORE FOR MALARIA 1% INCEPTION V3 WITH ATTACK.

Supervised			
class	precision	recall	f1-score
uninf.	0.55	0.72	0.62
parasit.	0.59	0.41	0.49
mean/std	0.57 ± 0.02	0.56 ± 0.21	0.55 ± 0.09
accuracy		0.569	
Semi-Supervised			
class	precision	recall	f1-score
uninf.	0.54	0.47	0.50
parasit.	0.53	0.60	0.56
mean/std	0.53 ± 0.007	0.53 ± 0.09	0.53 ± 0.04
accuracy		0.536	

the supervised scenario, however the semi-supervised network sees more (replicated) labeled instances per epoch.

V. Discussion

Overall, the semi-supervised network was able to improve results over the supervised version. A remarkable result is that, even CNN backbones that were not able to converge using only supervised learning: MobileNetV2 with 1% data, were significantly improved by using the unlabeled data, for example see Table IV in which the results jumped from random (10%) to almost 80%. More than that, for natural images (STL-10) and a more well-behaved dataset (Fashion), there seems to be even higher gains in attack scenarios when more labeled data is used (5%). It is expected that this improvement is less intense for (1%) since the attack relies mainly on the classification loss.

For the biomedical images domain, there are two main important observations: first because the images of Malaria are not angle-oriented, the rotation task becomes harder. Nevertheless, significant improvement was found when incorporating unlabeled data. Second, the attack had an even stronger impact on results due to the nature of the images, that contain patterns that are similar to the attack, degrading the results, as shown in two architectures in table VI. With only 137 images (1%) in the training set, it was possible to go from near random results in supervised learning up to 92% accuracy with our semi-supervised network (see Table V) which is comparable even when we use 5% of data that attained at most 94%.

VI. Conclusion

Self-supervised learning demonstrates to be useful in semi-supervised scenarios, not only to improve the numerical results, but in particular to learn more discriminative spaces, as well as a representation that is more robust with respect to attack. Our results showed that the choice of the auxiliary task must take into account the nature of the images and may not suit all applications. However, the mere introduction of unlabeled data into the training process significantly improved all minimal-data learning cases, e.g. using 1% of the available labels, even for a simple rotation prediction auxiliary task. This result may pose significant impact on applications in which annotation is costly, such as biomedical images.

As in previous and recent work, self-supervision appears as a relevant method to allow learning from minimal annotated data. More than that, when it is used during the learning process, it may help improving robustness against attacks. Future work may investigate other types of auxiliary tasks in the context of semi-supervised learning, as well as robustness against other undesired scenarios.

Acknowledgment

The authors would like to FAPESP (grants #2019/07316-0 and #2019/02033-0) and CNPq (National Council of Technological and Scientific Development) grant 304266/2020-5.

References

[1] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 2014, pp. 512–519. 1

[2] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2661–2671. 1

[3] M. A. Ponti, L. S. F. Ribeiro, T. S. Nazare, T. Bui, and J. Collomosse, "Everything you wanted to know about deep learning for computer vision but were afraid to ask," in *30th SIBGRAPI conference on graphics, patterns and images tutorials*. IEEE, 2017, pp. 17–41. 1

[4] G. B. Cavallari, L. S. Ribeiro, and M. A. Ponti, "Unsupervised representation learning using convolutional and stacked auto-encoders: a domain and cross-domain feature space analysis," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2018, pp. 440–446. 1

[5] F. P. Dos Santos, C. Zor, J. Kittler, and M. A. Ponti, "Learning image features with fewer labels using a semi-supervised deep convolutional network," *Neural Networks*, vol. 132, pp. 131–143, 2020. 1

[6] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1920–1929. 1

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738. 1, 2

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conf. on Machine Learning*. PMLR, 2020, pp. 1597–1607. 1, 2

[9] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[10] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020. 1, 2

[11] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009. 1

[12] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1195–1204. 1, 2

[13] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 1, 2

[14] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018. 1, 2

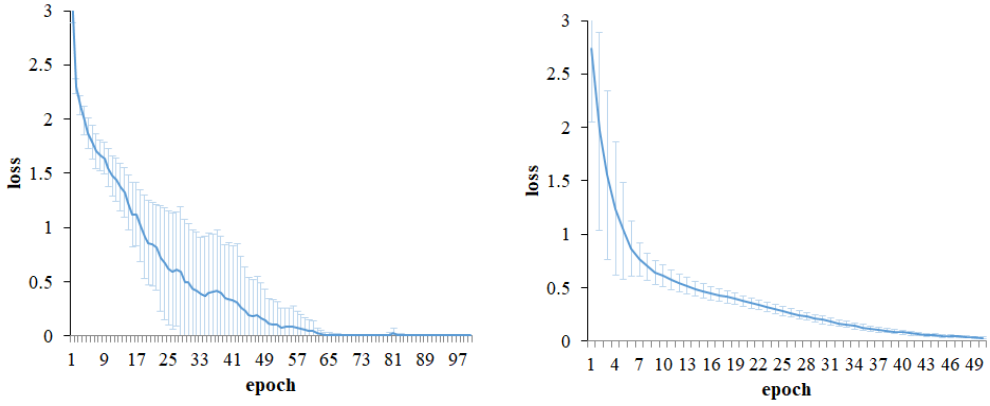
[15] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020. 1

[16] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *International Conference on Learning Representations*, 2020. 1

[17] C.-W. Kuo, C.-Y. Ma, J.-B. Huang, and Z. Kira, "Featmatch: Feature-based augmentation for semi-supervised learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 479–495. 1

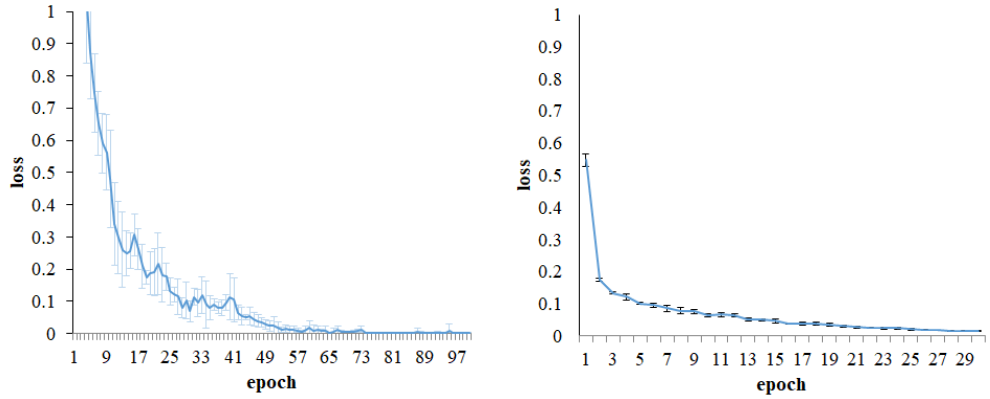
[18] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, 2018, p. 3239–3250. 1

[19] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *Proceedings of the International Conference on Learning Representations*, 2019. 1



(a) supervised (b) semi-supervised

Fig. 10. Training loss of models trained on STL-10 1% Resnet50 v2 with attack. (a) loss of the supervised CNN model, (b) loss of the combined semi-supervised model, requiring less epochs for convergence.



(a) supervised (b) semi-supervised

Fig. 11. Training loss of models trained on Fashion-MNIST 5% Mobilenet v2 without attack. (a) loss of the supervised CNN model, (b) loss of the combined semi-supervised model, requiring less epochs for convergence.

[20] J. R. Layza, H. Pedrini, and R. da Silva Torres, "1-to-n large margin classifier," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2020, pp. 316–323. 1

[21] B. Shi, J. Hoffman, K. Saenko, T. Darrell, and H. Xu, "Auxiliary task reweighting for minimum-data learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020. 2

[22] Y. Asano, C. Rupprecht, and A. Vedaldi, "A critical analysis of self-supervision, or what we can learn from a single image," in *International Conference on Learning Representations*, 2019. 2

[23] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430. 2

[24] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84. 2

[25] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9359–9367. 2

[26] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conf. on Computer Vision*, 2016, pp. 649–666. 2

[27] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations*, 2018. 2

[28] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1734–1747, 2015. 2

[29] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5898–5906. 2

[30] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 132–149. 2

[31] C.-Y. Chuang, J. Robinson, L. Yen-Chen, A. Torralba, and S. Jegelka, "Debiased contrastive learning," in *NeurIPS*, 2020. 2

[32] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1476–1485. 2

[33] P. V. Tran, "Exploring self-supervised regularization for supervised and semi-supervised learning," *arXiv preprint arXiv:1906.10343*, 2019. 2

[34] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Advances in Neural Information Processing Systems*, 2019. 2

[35] A. Coates, A. Ng, and H. Lee, "An Analysis of Single Layer Networks in Unsupervised Feature Learning," in *AISTATS*, 2011. 3

[36] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 3

[37] S. Rajaraman, S. K. Antani, M. Poostchi, K. Silamut, M. A. Hossain, R. J. Maude, S. Jaeger, and G. R. Thoma, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, p. e4568, 2018. 3

[38] R. F. Mello and M. A. Ponti, *Machine learning: a practical approach on the statistical learning theory*. Springer, 2018. 3